Análise de Componentes Principais

TUANY DE PAULA CASTRO

Finalidades e Aplicações

- **Objetivo:** Transformar um conjunto de variáveis originais em outro conjunto de variáveis da mesma dimensão denominadas componentes principais, retendo o máximo de informação em termos da variação total contida nos dados.

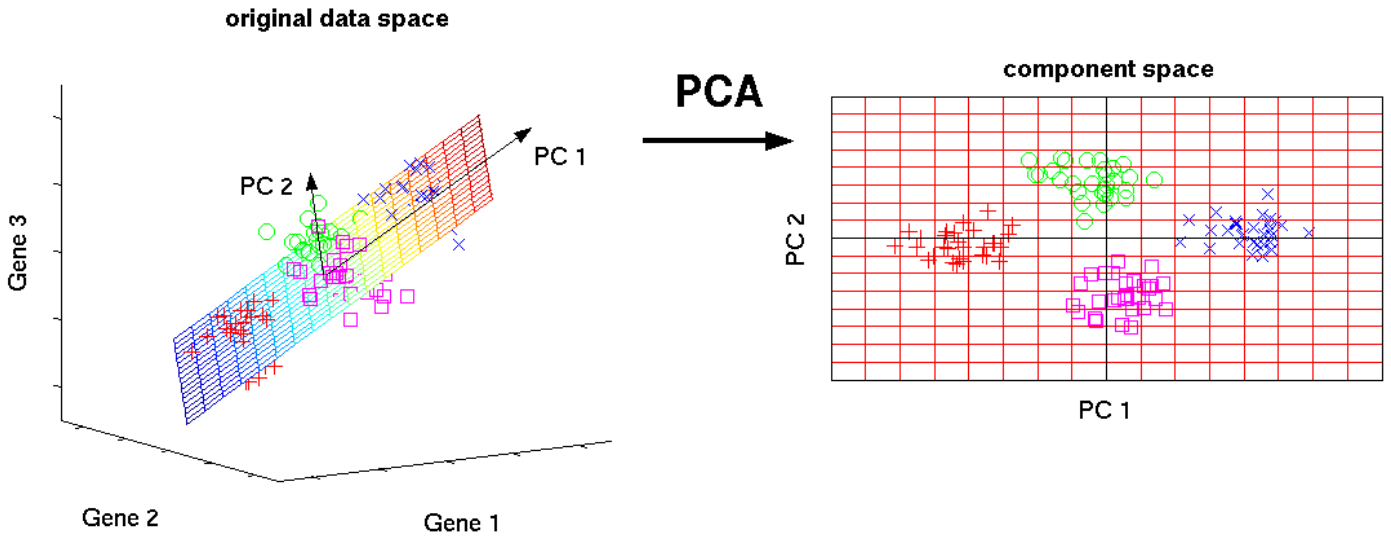
Quando usar?

- Quando se tem interesse em reduzir uma massa de dados garantindo a menor perda possível da informação.

- A análise de Componentes Principais frequentemente serve de etapa intermediária em investigações maiores, como entradas numa regressão múltipla ou análise de agrupamento ou ainda como método de extração de fatores na análise fatorial.

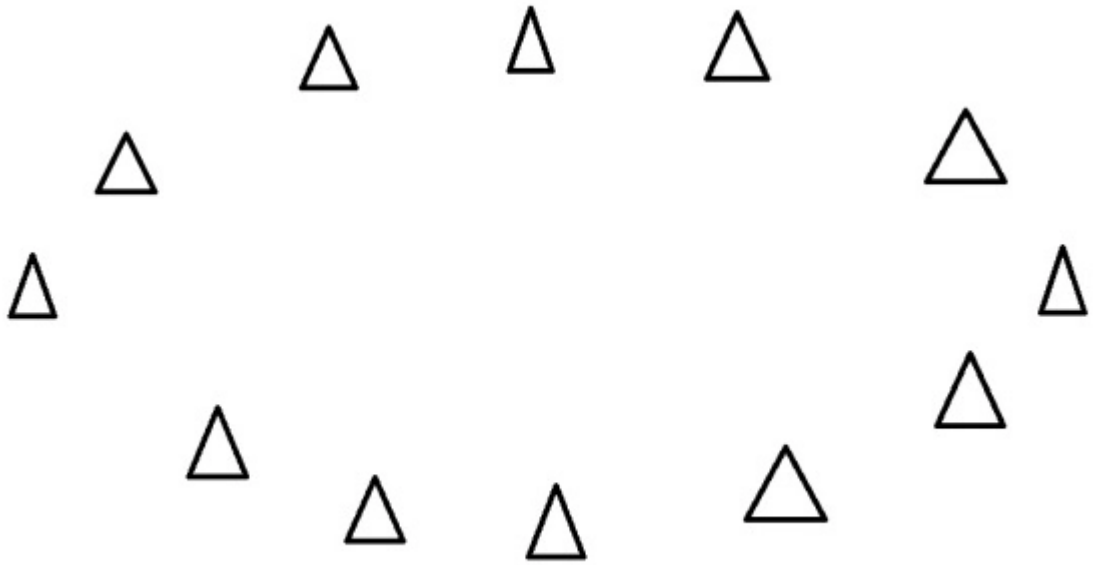
Obtenção das componentes

- ✓ Algebricamente, as componentes principais são combinações lineares das variáveis originais;
- ✓ Geometricamente, as componentes principais são rotações das variáveis originais representando o máximo de variabilidade.



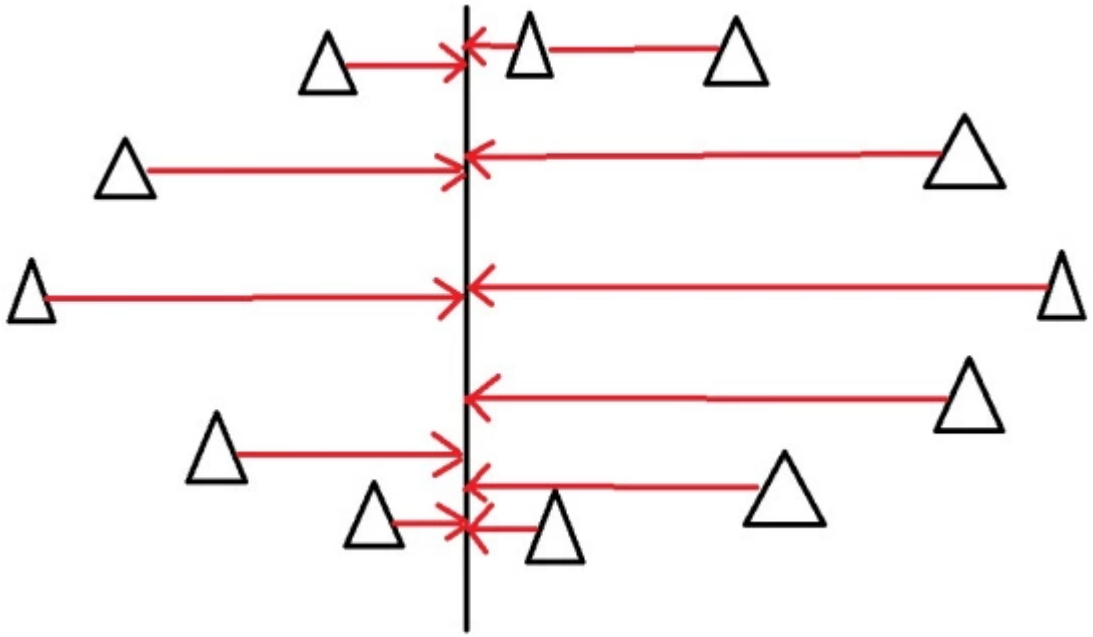
Obtenção das componentes

- Imagine que esses triângulos são observações de um banco de dados. Qual seria uma componente principal desses dados?



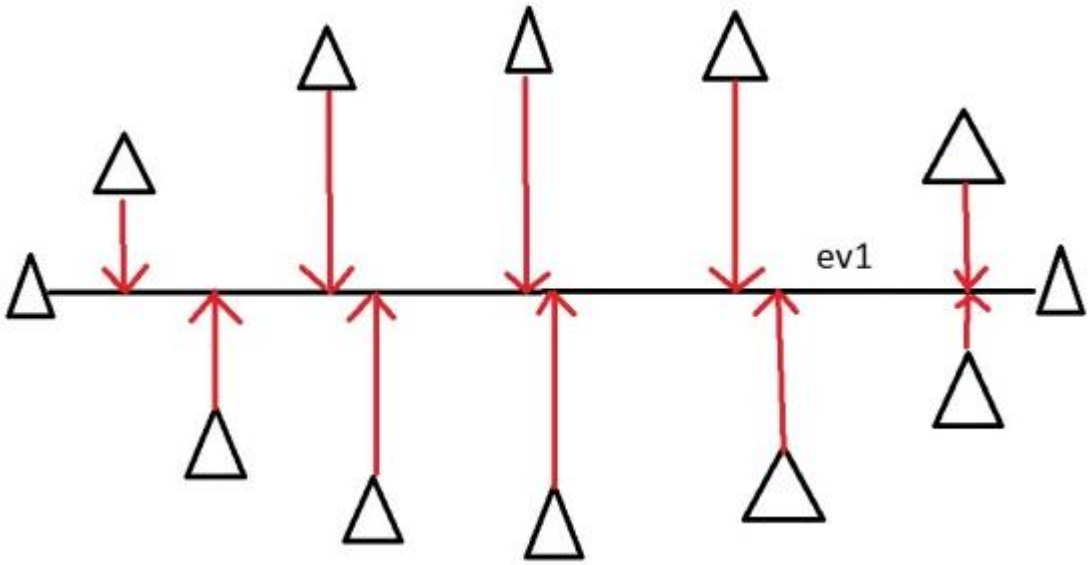
Obtenção das componentes

- As componentes principais são as direções onde há a maior variabilidade, onde os dados estão mais espalhados. Poderíamos tentar uma reta vertical:



Obtenção das componentes

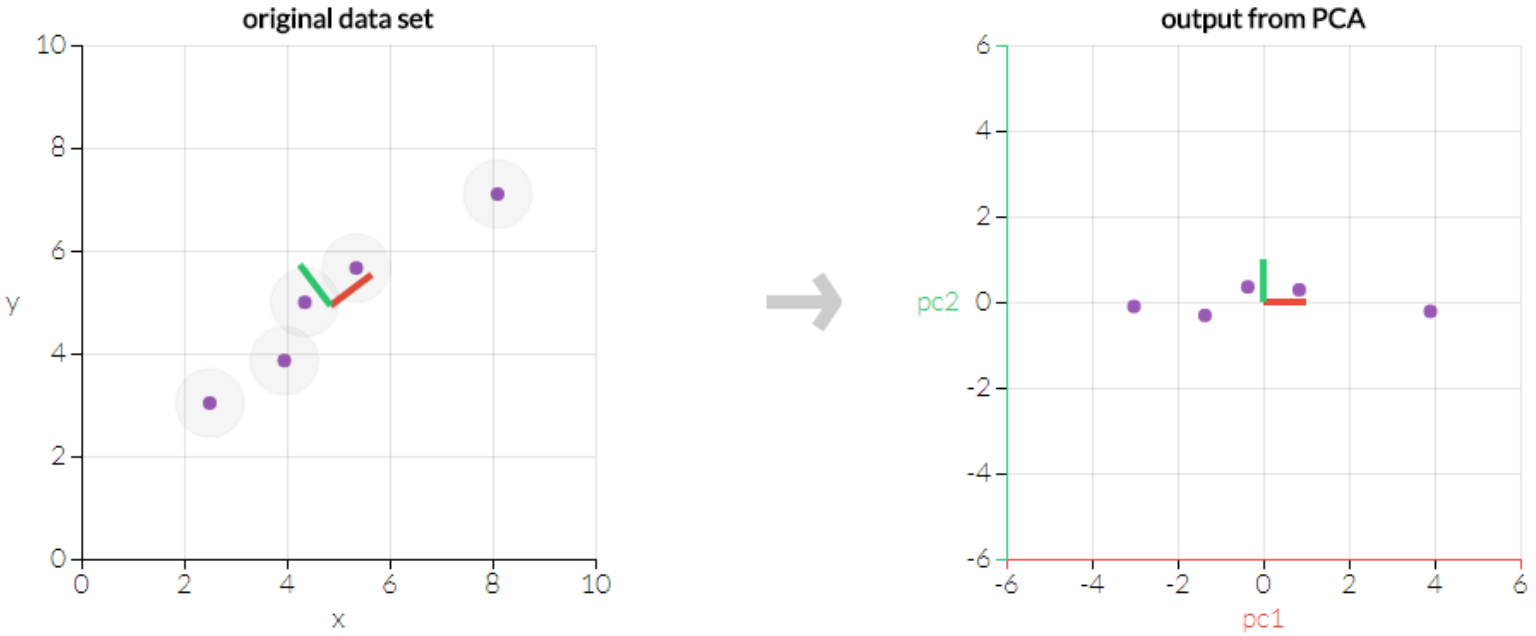
- Ou ainda uma reta horizontal:



- Em qual delas os dados ficam mais espalhados?

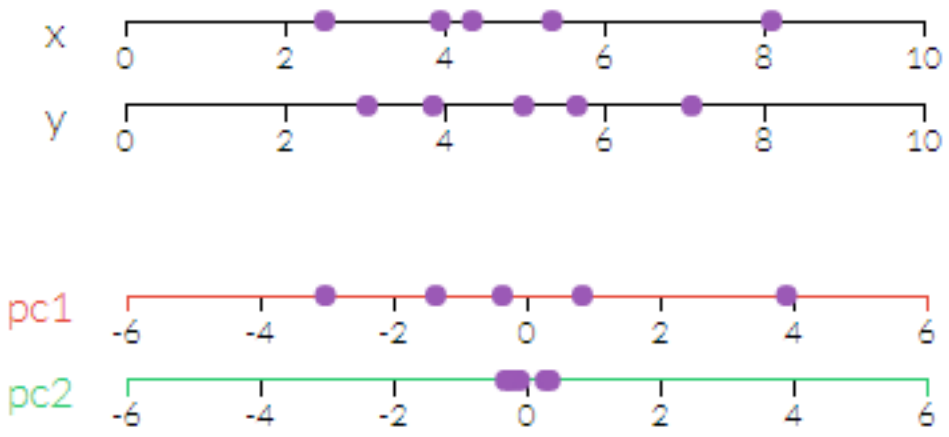
Obtenção das componentes

Em um banco de dados com duas variáveis (x e y):



Obtenção das componentes

Observando esses dados nos pares de linhas, vemos que não há grande perda se desconsiderarmos a segunda componente, pois a maior parte da variabilidade já é explicada com a primeira componente.



Obtenção das componentes

As componentes principais têm as seguintes propriedades:

- São **ordenadas** de maneira que a primeira componente tem a maior variância, a segunda tem a segunda maior variância, a terceira tem a terceira maior variância e assim sucessivamente;
- São **independentes**, ou seja, cada par de componentes tem correlação (covariância) nula.

Obtenção das componentes

As componentes principais podem ser obtidas por dois caminhos:

- **Matriz de covariância:** nesse caso, os resultados da análise são afetados pelas diferenças de variabilidade das variáveis originais. Indicado somente se a detecção de tais diferenças tiverem relevância para o estudo.
- **Matriz de correlação:** quando as diferenças de variabilidade se devem unicamente à diferença de escala das variáveis, incluir essa diferença na análise é desnecessária e uma alternativa é transformar os dados para remover esse efeito. A matriz de correlação é a matriz de covariância dos dados padronizados.

Escolha das componentes

A escolha da quantidade de componentes principais a serem analisadas pode ser feita considerando a contribuição das componentes na análise.

- A variância de cada componente é o seu autovalor;
- A contribuição de uma componente é a proporção da variância total explicada por ela:

$$\frac{autovalor\ da\ componente}{soma\ de\ todos\ os\ autovalores}$$

Escolha das componentes

Não há um critério consensual, porém há alguns métodos para auxiliar na definição:

Critério de Kaiser

✓ Devem ser analisadas apenas as componentes com autovalores acima de um.

Scree plot

✓ Analisar graficamente a dispersão do número de componentes até que a curva da variância individual de cada uma se torne horizontal ou sofra uma queda abrupta. Em ambas as situações, isso indica que variância foi perdida e, por isso, deve-se parar de extrair componentes.

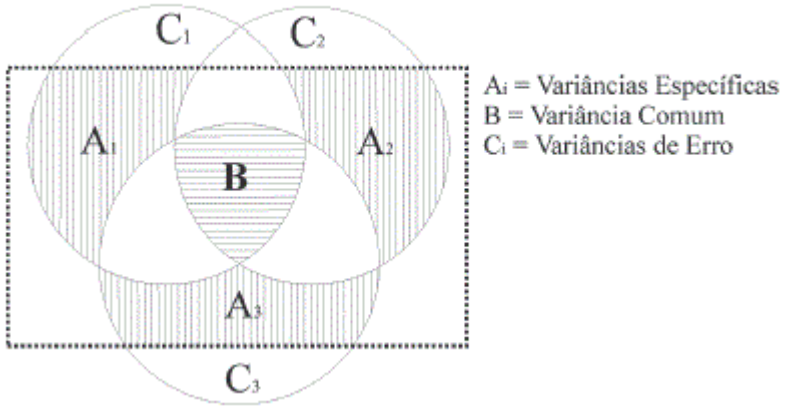
Critério da variância acumulada

✓ O mínimo depende do objetivo de uso da técnica. Recomenda-se pelo menos 70%, porém, se forem feitas outras análises, é adequado pelo menos 80%.

Diferença entre ACP e AF

Os métodos de redução de dados assumem que a variância de um fator ou componente é composto por três aspectos:

- **Variância específica:** variância de uma variável, não compartilhada com as demais;
- **Variância comum:** compartilhada entre todos os itens que compõem o fator ou componente;
- **Variância de erro:** parcela da variância não explicada pelo componente ou fator.



Diferença entre ACP e AF

Diferença entre Análise Fatorial e Análise de Componentes Principais:

	Análise de Componentes Principais	Análise Fatorial
Variância	Não diferencia as variâncias comum e específica	Considera apenas a variância comum.
Índices	Representam a variância comum e a específica.	Representam apenas a variância comum.
Vantagens	Variabilidade explicada mais elevada.	Mais precisão na compreensão dos fatores .

Diferença entre ACP e AF

Diferença entre Análise Fatorial e Análise de Componentes Principais:

" Se você estiver interessado numa solução teórica não contaminada por variabilidade de erro, a análise fatorial deve ser sua escolha. Se você quiser simplesmente um resumo empírico do conjunto de dados, a análise de componentes principais é uma escolha melhor." (Tabachinick e Fidell, 2007)

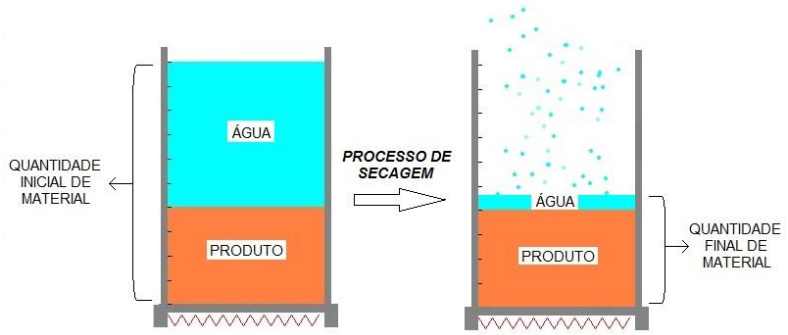
"A análise de componentes principais é em geral preferida para fins de redução de dados, enquanto a análise fatorial é em geral preferida quando o objetivo da pesquisa é detectar a estrutura dos dados da modelagem casual" (Garson, 2009)

Estudo das relações entre variáveis

A ACP também pode ajudar a analisar a relação entre as variáveis originais (a maneira como se correlacionam e como influenciam na determinação do novo sistema de coordenadas). As quantidades a serem analisadas são as cargas, indicando alta influência quando próximas de 1 ou -1.

Exemplo

Considere o arquivo *dados-evaporador-industrial.csv*. Esses dados contém detalhes do processo de secagem de uma quantidade fixa de um produto úmido colocado em um leito evaporador. Durante o processo de evaporação, foram avaliadas 8 variáveis em intervalos de tempo regulares para monitoramento e controle de qualidade: ponto de orvalho, temperatura de entrada, temperatura do ar em processo, temperatura de exaustão, fluxo de massa do ar, temperatura da cama, pressão do filtro e pressão da cama. Vamos aplicar a técnica de Componentes Principais para redução dos dados.



Exercícios

1) Considere os *dados-poluição.xlsx*. Resuma esses dados em menos de 7 componentes principais usando a matriz de covariância e a matriz de correlação. O que você observa? Há alguma diferença nos resultados pela escolha da matriz? Esses dados podem ser resumidos em 3 ou menos dimensões? Pode-se interpretar as componentes obtidas?

2) Considere o conjunto *dados-passaros.xlsx* com observações de 49 pássaros quanto ao comprimento total (X1), extensão alar (X2), comprimento do bico e cabeça (X3), comprimento do úmero (X4) e comprimento da quilha do esterno (X5). Faça uma análise de componentes principais para reduzir o número de variáveis. Interprete e comente os resultados.

Exercícios

3) Em *dados_mc.csv,* encontram-se dados referentes a uma avaliação nutricional de produtos vendidos pela rede MC Donald's.

(a) Identifique e interprete as componentes principais que caracterizam os produtos nutricionalmente.

(b) Compare as categorias de produtos com relação a essas componentes. Há diferença significativa entre as categorias? Identifique essas diferenças por meio de um modelo de regressão.

Exercícios

4) Em diabetes.csv, encontram-se os dados referentes a 768 mulheres avaliadas com relação a medidas de diagnóstico e à diabetes. As variáveis medidas são: número de vezes em que engravidou (Pregnancies), concentração de glicose no plasma (Glucose), pressão arterial diastólica (BloodPressure, em mmHg), espessura da pela do tríceps (Skin Thickness, em mm), insulina (Insulin, em mm U/ml), índice de massa corporal (BMI, em kg/m^2), função de tipo da diabetes (Diabetes Pedigree Function), idade (Age, em anos) e presença de diabetes (Outcome: 0 se não tem e 1 se tem).

(a) Faça uma análise de componentes principais para reduzir a dimensão dos dados, desconsiderando a variável Outcome. Comente e tente interpretar os resultados.

(b) Utilize os resultados da ACP para construir um modelo de previsão para a presença de diabetes. Comente.

Referências

- BISQUERRA, R; CASTELLA, J.; VILLEGAS, F. Introdução à estatística: enfoque informático com o pacote estatístico SPSS. Porto Alegre: Artmed, 2007.
- DANCEY, Cristine P; REIDY, John. Estatística sem Matemática Para Psicologia. 3 edição. Porto Alegre: Artmed, 2006.
- HAIR, J.; ANDERSON, R.; BLACK, W. Análise multivariada de dados. 6 ed. Reimp. Porto Alegre: Bookman, 2009.
- JOHNSON, R. and WICHERN, D. Applied Multivariate Statistical Analysis. Sixth edition, Wisconsin, Pearson, 2007.